Effect of lecture instruction on student performance on qualitative questions

Paula R. L. Heron*

Department of Physics, Box 351560, University of Washington, Seattle Washington, 98195-1560, USA

(Received 13 May 2014; published 22 January 2015)

The impact of lecture instruction on student conceptual understanding in physics has been the subject of research for several decades. Most studies have reported disappointingly small improvements in student performance on conceptual questions despite direct instruction on the relevant topics. These results have spurred a number of attempts to improve learning in physics courses through new curricula and instructional techniques. This paper contributes to the research base through a retrospective analysis of 20 randomly selected qualitative questions on topics in kinematics, dynamics, electrostatics, waves, and physical optics that have been given in introductory calculus-based physics at the University of Washington over a period of 15 years. In some classes, questions were administered after relevant lecture instruction had been completed; in others, it had yet to begin. Simple statistical tests indicate that the average performance of the “after lecture” classes was significantly better than that of the “before lecture” classes for 11 questions, significantly worse for two questions, and indistinguishable for the remaining seven. However, the classes had not been randomly assigned to be tested before or after lecture instruction. Multiple linear regression was therefore conducted with variables (such as class size) that could plausibly lead to systematic differences in performance and thus obscure (or artificially enhance) the effect of lecture instruction. The regression models support the results of the simple tests for all but four questions. In those cases, the effect of lecture instruction was reduced to a nonsignificant level, or increased to a significant, negative level when other variables were considered. Thus the results provide robust evidence that instruction in lecture can increase student ability to give correct answers to conceptual questions but does not necessarily do so; in some cases it can even lead to a decrease.

DOI: [10.1103/PhysRevSTPER.11.010102](https://doi.org/10.1103/PhysRevSTPER.11.010102)

PACS numbers: 01.40.Fk

I. INTRODUCTION

The impact of lecture instruction on student conceptual understanding in physics has been the subject of research for several decades. Most studies have reported disappointingly small improvements in student performance on conceptual questions despite direct instruction on the relevant topics [1]. These results have spurred a number of attempts to improve learning in physics courses through new curricula and instructional techniques [2].

This paper contributes to the research base through a retrospective analysis of results from 20 randomly selected qualitative questions on kinematics, dynamics, electrostatics, waves, and physical optics that have been given in introductory calculus-based physics at the University of Washington (UW) over a period of 15 years. The questions were administered in a large number of classes, in some of which relevant lecture instruction had been completed, and in some of which it had yet to begin. For some questions there was a statistically significant difference in average

performance before and after lecture instruction; in others, the results were indistinguishable. However, because the classes had not been randomly assigned to be tested before or after lecture instruction, there are many other variables, such as class size, that might be important. The analysis presented here attempts to determine whether the results of simple comparisons can be taken at face value; that is, whether an observed difference can properly be attributed to lecture instruction and, in those cases where no difference is observed, whether the effects of lecture instruction are absent or simply hidden.

II. ABOUT THE DATA

The data analyzed in this paper were obtained in introductory calculus-based physics at UW, which is a sequence of three courses: Mechanics; Electricity & Magnetism; and Waves & Optics. Each course is offered every 10-week academic quarter in multiple lecture sections of up to 225 students each. For example, in autumn quarter there are typically three sections of Mechanics, three of E&M, and one of Waves & Optics. Lectures are usually taught by faculty members in the physics department. A standard textbook is used, from which weekly homework sets are assigned. Since 2000, homework has been administered and graded online. In addition to three weekly 50-min lectures, students attend weekly labs,

* pheron@uw.edu

Published by the American Physical Society under the terms of the Creative Commons Attribution 3.0 License. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

which are taught by graduate and undergraduate teaching assistants.

This course is also the context in which the UW Physics Education Group has been developing *Tutorials in Introductory Physics*, a set of materials intended to supplement instruction in typical introductory courses [3]. During weekly tutorials, students work in small groups on sequences of experiments and exercises designed to help them construct concepts and develop reasoning ability. Prior to each tutorial, students complete a “pretest,” from which the data in this study derive.

Pretests, which consist of a set of conceptual questions, serve several purposes. For students, they highlight the issues that will be addressed in the coming tutorial and provide practice on the types of questions that will appear on course examinations (all exams contain questions on lecture, lab, and tutorial content). Pretest responses also inform teaching assistants about their students’ ideas, and about the common difficulties that may emerge during the tutorial. For tutorial developers, quantitative and qualitative analysis of pretest responses provides insight into student understanding of topics covered in the course, and reveals difficulties to be addressed. Finally, pretest performance provides a benchmark against which to measure post-tutorial performance.

It is important to emphasize that “pretests,” so named because they precede each week’s tutorial, are intended to be administered *after* lecture instruction on the topics addressed in the corresponding tutorial. For example, the tutorial *Newton’s Second and Third Laws*, and its accompanying pretest, are intended to follow lecture instruction on Newtonian dynamics and address conceptual difficulties that often linger. However, due to occasional changes in the course schedule, the pretest is sometimes given after lecture instruction on forces has begun, but before formal introduction of all of Newton’s laws. In other quarters, most of the unit on Newton’s laws, including homework and laboratory experiments, has been completed. Thus pretests are administered in one of two conditions: *lecture-pretest-tutorial* or *pretest-lecture-tutorial*. Data collected in the first of these conditions will be referred to throughout this paper as “after lecture;” data collected in the second, as “before lecture.” In either case the pretest occurs *before* the relevant tutorial, the impact of which is not directly examined here.

Since 2000, the tutorial pretests have been administered online. Students have 15 min to take the pretest, which they can do at any time during a roughly 72-h period that starts after their Friday lecture and ends before their Monday lecture. Several conceptual questions are posed. Students are asked to select answers from a menu of choices and to type brief explanations in a text box. They receive a small amount of credit, whether or not their answers are correct. Periodic spot checking helps ensure that students take the pretests seriously. Up to 1400 students take the introductory

calculus-based course each quarter and some pretests have been given many times over the past two decades. Thus the responses constitute a large data set.

A. Selection of pretest questions for analysis

The first step in the study reported here involved identifying all pretest questions that (a) had been given in a minimum of 20 lecture sections (to screen out questions for which samples would almost certainly be too small); and (b) could be easily analyzed (e.g., the questions asked for a ranking, a before-and-after comparison, etc.). A random selection of 30 questions (in sets of one to four) was then chosen for further analysis, in order to avoid bias in favor of a certain type of question or topic area. Questions that did not involve enough instances of both instructional conditions were dropped. Another was dropped because the success rate even before instruction was above 95%. The final list of 20 questions can be found in Table I.

B. Selection of classes

With a few exceptions, all classes for which data were available were included. Honors sections of the introductory course were excluded because performance in those classes is typically higher than in nonhonors sections. Also excluded were summer quarter classes. Many students who enroll in the summer are enrolled in degree programs elsewhere, are catching up, or retaking classes.

Finally, all classes in which fewer than 50% of the students took the relevant pretest were excluded. When this occurs it is usually the result of a technical problem with the online system on which the pretests are administered. Therefore results in these sections are not a reliable indicator of what performance would have been had no problem occurred.

III. OVERVIEW OF ANALYSIS

The analysis described in this paper represents an attempt to determine, for each question, if lecture instruction has an impact on average performance. The analysis had two stages. Stage 1 involved simple tests of statistical significance to identify cases in which a difference in performance exists, and those in which it does not. Stage 2 involved multiple linear regression to understand better the reasons a difference might, or might not, appear. Specifically, the second stage attempts to assess whether results that appear to indicate that lecture instruction affects performance are better explained by some other variable. Similarly, for cases that appear to indicate that lecture instruction does *not* affect performance, the second stage serves to assess whether the effect is merely hidden by other variables.

Below, the designation of classes as after lecture or before lecture is described and performance is defined more

precisely. Subsequent sections describe stages 1 and 2 in detail.

A. Designating classes as before lecture or after lecture

The data analyzed in this study were obtained over a 15-yr period. There are no records that indicate what happened in every lecture period for every class, only what was scheduled to occur. Therefore, the decision as to whether a given class belongs in the before lecture or after lecture sample depends only on whether or not relevant lecture instruction had been scheduled to occur before the questions became available to students. Efforts to mitigate the effects of possible errors in categorization are discussed later.

The decision as to what constitutes “relevant instruction” differs for different questions. Questions 12 to 15 in Table I, drawn from the pretest for the *Capacitance* tutorial, provide an illustration. Each question deals with a pair of conducting plates connected to an ideal battery. The questions concern changes in the potential difference (No. 12), electric field (No. 13), charge densities (No. 14), and capacitance (No. 15) that would occur if the plates were moved closer together. For the first three, relevant instruction would include the electric field between “infinite” charged plates, the relationship between potential difference and electric field, and the role of an ideal battery in maintaining a constant potential difference. For the fourth, the minimum relevant instruction would also include the definition of capacitance; instruction on parallel-plate capacitors would be relevant but not necessary. Thus a class could be in the after lecture sample for questions 12 to 14, but in the before sample for question 15.

B. Defining performance

The *performance* on question i of class j , π_{ij} is defined as

$$\pi_{ij} = \frac{N_{ij}^{\text{corr}}}{N_{ij}^{\text{total}}},$$

where N_{ij}^{corr} and N_{ij}^{total} are the number of correct answers and the total number of answers received, respectively. (In general, N_{ij}^{total} is less than the total enrollment in class j , N_j .) Performance thus ranges from 0 to 1 and is quoted throughout this paper to two decimal places.

It is important to note that performance, as defined here, is a very crude measure of understanding. Examining the explanations offered by students can help gauge whether correct answers stem from solid understanding of the topics, and whether incorrect responses reveal deficiencies. However, the explanations given by students on pretests are often very brief and consequently difficult to assess. Moreover, the scale of the study described here, which

involved more than 120 000 individual student answers, necessitated a simple, objective measure.

The average performance on question i of a sample of classes π_i^{ave} is

$$\pi_i^{\text{ave}} = \frac{1}{N_i} \sum_j \pi_{ij},$$

where N_i is the number of classes in the before lecture sample or the after lecture sample.

Averages (and other summary statistics) reported in this paper give all classes equal weight, regardless of the number of students. Weighted averages (and weighted standard deviations) were calculated but are not quoted here, in part because the unweighted numbers provide a more direct basis for comparison with some of the statistical tests reported below, and with the results of the regression analysis. In the majority of cases, the difference between the weighted and unweighted averages was less than 0.01; the maximum was 0.02. The average class size is about 150 so a difference of 0.01 represents fewer than two students. Still, such a difference could, in principle, be important. However, unless explicitly mentioned, the results of significance testing were essentially the same for weighted and unweighted quantities, at least insofar as establishing a level of significance.

IV. STAGE 1: EXAMINING DATA FOR DIFFERENCES BETWEEN PERFORMANCE OF BEFORE LECTURE AND AFTER LECTURE SAMPLES

In the sample scatter plots in Fig. 1, performance is plotted for the entire dataset (black diamonds), the classes in the before lecture sample (gray diamonds), and the classes in the after lecture sample (white diamonds). As shown, in all but one case the before and after sets overlap, such that the performance of some prelecture classes is superior to that of some postlecture classes. Nevertheless, in a few cases there seems to be a clear tendency for performance to differ.

A. Explanation of statistical tests and their interpretation

Tests of significance were conducted to calculate the probability that an observed difference would occur by chance if the two samples were drawn at random from the same population, where *population* refers to the set of all introductory calculus-based physics classes at UW, not to a population of students. Because some samples are small it is not possible to determine whether the conditions for applicability of certain statistical tests are satisfied in every case (e.g., that the distributions being compared are normal, as assumed for a t test). Two strategies were adopted to mitigate the possible effects: triangulating by using more

TABLE I. Selected questions (summarized).

Topic area	Scenario	No.	Question
Acceleration in 1D	A ball rolls up and down an incline.	1	What is the direction of the acceleration of the ball at the top of the incline? (Choose an arrow or “zero.”)
		2	Does the magnitude of the acceleration increase, decrease or remain the same as the ball rolls up the incline?
		3	Does the magnitude of the acceleration increase, decrease or remain the same as the ball rolls down the incline?
Newton’s 3rd Law	Two pucks collide; velocity vectors are shown before and after the collision	4	Is the magnitude of the acceleration of puck <i>A</i> greater than, less than, or equal to that of puck <i>B</i> ?
	A crate is on a platform; both move upward at constant speed	5	Is the magnitude of the force on the crate by the platform greater than, less than, or equal to that of the force on the platform by the crate?
	A crate is on a platform; both move upward with increasing speed	6	Is the magnitude of the force on the crate by the platform greater than, less than, or equal to that of the force on the platform by the crate?
Dynamics of rigid bodies	Two identical spools are connected by a thread that passes over a frictionless pulley. One end of the thread is wound around one spool; the other end is tied to the other spool. The spools are released from the same height at the same instant.	7	Is the acceleration of the center of mass of spool <i>A</i> greater than, less than, or equal to that of spool <i>B</i> ?
		8	Which spool will hit the ground first?
Area vectors	Two sheets of paper are shown.	9	Choose (from a series of drawings) the arrows that represent the orientations and (relative) sizes of the sheets of paper.
Flux	A closed imaginary surface is shown; the surface is in a uniform electric field.	10	Is the electric flux through the entire surface positive, negative, or zero?
Electric fields, potential and capacitance	Two isolated identical plates have equal and opposite charge densities. They are moved closer together.	11	Does the capacitance of the arrangement increase, decrease, or remain the same?
		12	Does the potential difference between the plates increase, decrease, or remain the same?
	Two identical plates are connected by an ideal battery. They are moved closer together.	13	Does the electric field between the plates increase, decrease, or remain the same?
		14	Does the charge density on each plate increase, decrease, or remain the same?
		15	Does the capacitance of the arrangement increase, decrease, or remain the same?
Reflection & transmission of pulses	Two springs are joined at a junction; pulses travel faster on the left spring. A pulse travels toward the junction from the left.	16	Choose (from a series of drawings) the shape of the left spring a short time after the pulse has reached the junction
		17	Choose (from a series of drawings) the shape of the right spring a short time after the pulse has reached the junction
Two source interference	Two in-phase sources tap the surface of a tank of water; one point is marked that is on a line that passes through both sources.	18	Is the point a point of maximum constructive interference, complete destructive interference, or neither?
Single slit diffraction	A diffraction pattern produced by laser light of wavelength λ incident on a slit.	19	Is the slit width greater than, less than, or equal to λ ?
		20	If half the slit were covered, would the width of the central maximum of the pattern increase, decrease or remain the same?

than one type of test, and adopting a somewhat loose interpretation of statistical significance.

To triangulate, both one-tailed *t* tests (actually a variant designed for data that may violate the assumption of equal variance) and nonparametric *permutation tests* were conducted [4]. The latter type of test has fewer requirements.

The permutation test is based on shuffling the data points and their category labels, effectively creating fictional data sets. Under the null hypothesis, that the two samples are drawn from the same population, the real data set will be indistinguishable from the fictional ones. The first question in Table I provides an illustration. Of a total of 79 data

FIG. 1. Scatter plots for representative questions (5, 6, 8, 10, 13, and 20 in Table I). Each diamond represents a single class. Black diamonds represent all classes; gray and white diamonds represent classes questioned before lecture instruction and after lecture instruction, respectively.

points (79 classes), 58 have the label *before lecture* and 21 have the label *after lecture*. These data could be shuffled to create $79!/(58!21!)$ distinct fictional data sets, in each of which 58 data points will carry the before lecture label while the rest carry the after lecture label. (If the number of combinations is sufficiently large, as it is here, a random sample of them is generated instead.) The difference of the means is calculated for each fictional data set. The percentage of fictional sets in which this difference exceeds that of the actual set can be interpreted as the probability of such a difference occurring by chance.

A somewhat loose interpretation of statistical significance also helps ensure that test results are not assumed to be more robust than warranted. Specifically, throughout this paper, a difference in means that is significant at the 10% level is taken as evidence of an effect due to lecture instruction.

This choice also has another implication. The appropriate level for declaring a result statistically significant depends partly on conventions in various disciplines, and

also on the plausibility of the claim, and the stakes involved in its acceptance or rejection. If an intervention were being tested, the null hypothesis would be that the intervention is not effective and any observed difference is in fact just due to chance. Strong evidence is generally needed to support rejection of this hypothesis. The more stringent the criterion for rejecting the null hypothesis (e.g., $p < 5\%$, 1% , or lower) the more heavily weighted is the decision in favor of the usual practice (or control condition). In contrast, the “intervention” here is simply lecture instruction. Thus the null hypothesis is that lecture instruction does *not* improve performance and any observed difference is just due to chance. This hypothesis runs counter to conventional wisdom and is difficult for many instructors to believe. The criterion for statistical significance adopted here acknowledges that even weak evidence is likely to be taken as grounds for rejecting this hypothesis. As a result, even a difference in performance that could happen almost 10% of the time by chance will be considered evidence of the impact of lecture instruction.

B. Results

The two types of tests were in agreement for 19 of the 20 questions examined, at least insofar as designating results significant at the 10%, 5%, or 1% level. The weighted and unweighted t tests were also in accord. However, for question 11, the p values were just above 1% (permutation test and weighted t test) and just below 1% (unweighted t test). In the discussion that follows, this result is assumed to be significant at the 5% level.

The results are displayed in Table II. The first several columns contain summary statistics: sample sizes, means, and standard deviations of the before lecture and after lecture samples. As can be seen, for some questions, average performance after instruction reached 0.70 (i.e., 70% of students answering correctly); while in others, average performance both before and after instruction was well below 0.50. The standard deviations range from 0.04 to 0.14, with 80% falling between 0.05 and 0.09.

As shown, the average performance of the after lecture classes was significantly better than that of the before lecture classes for 11 questions, significantly worse for two questions, and indistinguishable for the remaining seven [5]. The differences (after–before) ranged from -0.05 (No. 12) to $+0.37$ (No. 10).

Table II also includes standardized mean difference effect sizes (Cohen's d), which range in absolute value from 0.0 to 5.9, with a median just over 0.7. Effect sizes, unlike tests of significance, are not influenced by sample size. Instead, they provide a measure of the difference between (estimated) population means in terms of the (estimated) population standard deviations. For example, the largest effect size in Table II ($d = 5.9$ for question 10) suggests the means are nearly 6 standard deviations apart, an effect that should be (and is) easily visible to the naked eye (see Fig. 1). There are no generally accepted standards for designating effect sizes as “large,” “small,” etc., that transcend the type of measurement being conducted [6]. However, they incorporate more information than raw difference. The subject of *practically* important, or “instructionally significant,” effects is discussed later.

The results are also shown graphically in Fig. 2. The bars represent the raw differences in means, arranged in increasing order. The order is different than that in Table I or II, but the labels correspond. The bars are shaded according to the result of the significance testing described above, with the dark, medium, and light gray bars indicating significance at the 1%, 5%, and 10% levels, respectively. Not surprisingly, the larger differences were more likely to be deemed statistically significant.

TABLE II. Summary statistics, results from hypothesis testing and effect sizes.

	Before lecture instruction			After lecture instruction			Results from significance testing	
	No. of sections	Average performance	Standard deviation	No. of sections	Average performance	Standard deviation	Difference	Effect size
	N_{before}	π_{before}	σ_{before}	N_{after}	π_{after}	σ_{after}	$\pi_{\text{after}} - \pi_{\text{before}}$	d
1	58	0.27	0.08	21	0.28	0.05	0.01	0.1
2	58	0.44	0.09	21	0.51	0.08	0.07***	0.8
3	58	0.45	0.09	21	0.53	0.05	0.09***	1.1
4	29	0.56	0.05	13	0.56	0.04	0	0.0
5	17	0.69	0.09	37	0.69	0.11	0	0.0
6	17	0.21	0.06	37	0.30	0.07	0.08***	1.2
7	21	0.41	0.05	18	0.38	0.06	-0.04^{**}	0.8
8	21	0.37	0.05	18	0.36	0.05	-0.01	0.3
9	57	0.57	0.09	5	0.70	0.05	0.14***	1.6
10	41	0.18	0.06	5	0.55	0.08	0.37***	5.9
11	9	0.47	0.06	46	0.53	0.09	0.07**	0.8
12	13	0.40	0.09	39	0.35	0.08	-0.05^{**}	0.6
13	13	0.40	0.06	39	0.41	0.08	0.01	0.1
14	13	0.24	0.04	39	0.28	0.05	0.04***	0.9
15	9	0.37	0.05	43	0.48	0.09	0.12***	1.4
16	35	0.19	0.05	17	0.20	0.10	0.0	0.2
17	35	0.20	0.05	17	0.18	0.04	-0.02	0.3
18	45	0.49	0.09	7	0.55	0.08	0.05*	0.6
19	9	0.37	0.06	46	0.46	0.08	0.08***	1.1
20	9	0.46	0.11	46	0.56	0.14	0.1***	0.7

***, ** and * indicate significance at the 1%, 5%, and 10% levels, respectively.

FIG. 2. Raw differences in average performance arranged in increasing order. The labels correspond to those in Tables I and II. Bars shaded dark gray, medium gray, and light gray indicate effects significant at the 1%, 5%, and 10% levels, respectively.

V. STAGE 2: EXAMINING DATA FOR EFFECTS OF OTHER VARIABLES

The interpretation of the results reported above would be (comparatively) straightforward if the classes had been randomly assigned to be tested either before or after lecture instruction. In that case, the sample consisting of classes tested before lecture and that consisting of classes tested after lecture would have comparable numbers of classes from different quarters, different times of day, etc. While some variation from class to class in the background of the students would be expected, one could assume that neither *sample* was strongly biased in favor of higher preinstruction performance than the other. One could also assume that neither sample was strongly biased in favor of higher average motivation or ability, smaller class size, or other factors that could lead to greater improvements in performance even given identical instruction. Thus, although each *class* would have a somewhat different composition and experience somewhat different instruction, the average performance of both *samples* would be expected to be essentially the same if tested at the same stage. Therefore any difference in average performance between the samples could be attributed to the effect of lecture instruction. If no such difference were observed, it could be that no class experienced an effect due to lecture instruction, or that comparable numbers of classes experienced positive and negative effects.

However, the timing of the questions relative to lecture instruction in this study was *not* the result of randomization, but instead determined by the course schedule, which undergoes occasional changes due to changes in the topics covered, the order in which they are covered (usually due to adoption of a new textbook), or the timing of holidays in a given academic quarter. It is possible that these changes could also lead to systematic differences in some quality that would affect performance and thus render before lecture and after lecture samples noncomparable. For

example, instruction on Newton's laws is often scheduled to begin on the Friday of the second week of instruction in winter and spring quarters, but the Monday of the third week in autumn. The pretest scheduled for the intervening weekend would then occur before any lecture instruction in autumn, but after (some) lecture instruction otherwise. If Mechanics classes in autumn quarter preferentially enroll students of higher ability, motivation or preparation, then intrinsically higher performing classes might be more heavily represented in the before lecture sample than in the after lecture sample. If no difference in performance is seen, it could simply mean that the after classes started at a lower average level and lecture instruction raised them to the starting level of the before classes. Likewise, a difference in performance that *appears* to be due to lecture instruction might in fact be a spurious result of some other difference between the samples.

To gauge the degree to which systematic differences between the before lecture and after lecture samples are responsible for observed differences in performance (or lack thereof), multiple linear regression was conducted. Below the selected variables and the regression models are discussed.

A. Selected variables

As noted, while many class-level variables (e.g., class size, or most common major) could lead some classes to perform better than others, the present study considered only those that could cause a *sample of classes* to perform better than another. These are variables that might vary systematically with the timing of scheduled instruction on a given topic *and* that might predispose certain classes to have superior preinstruction performance, or to improve more given identical instruction. Variables that might lead to differences in performance among the after lecture classes, such as the textbook in use, the homework problems assigned, etc., are not relevant at this stage.

1. Variables associated with average student characteristics

The average preparation, motivation, or ability of students in a class can be expected to affect performance prior to instruction, as well as to affect the degree to which different classes respond to identical instruction. These characteristics might lead to systematic differences in performance between the before lecture and after lecture samples if those samples have different distributions of intrinsically higher and lower performing classes. This might occur if classes at certain times of year or at certain times of day preferentially enroll students with stronger background preparation or academic ability. For example, students in more or less competitive majors might have schedule conflicts that constrain their enrollment in introductory physics. Controlling directly for average student characteristics (e.g., SAT scores, planned major, highest level of high school physics, etc.) is one approach to mitigating this possibility [7]. However, because these will only affect the results if they are associated with schedule features, controlling for time of day and quarter effectively controls for variations in student characteristics.

Another variable that could cause before lecture and after lecture samples to differ in terms of average ability or preparation relates to the upward trend in average SAT scores of incoming UW students. Average SAT scores have increased steadily, from about 1150 for 1997–1998 incoming freshmen to 1212 for 2011–2012 incoming freshmen [8]. Therefore, it is plausible that classes in the more recent past include students predisposed to higher performance. Because schedules change every few years it is possible that these classes also tend to be more heavily represented in the before or after samples. Therefore, academic year is considered in the regression analysis.

2. Variables associated with the general conditions of lecture instruction

Regardless of the planned contents of a lecture, its impact could be influenced by the alertness of the instructor and/or the students. Therefore the time of day is, again, a potentially important variable. Greater communication among students and/or between students and the instructor might also occur in smaller classes and lead to greater improvement in performance. Therefore class size (defined as the number of students enrolled) is considered. A complication is that the level of attendance, which is not recorded, will clearly affect the number of students in a given class who were, in fact, exposed to instruction. In the past few years the use of clickers has provided a *de facto* attendance system, and it is possible that the opportunity to earn additional points has spurred greater class attendance, but no reliable data are available. However, attendance is only relevant if it is different, on average, for before and after samples. We have no reason to expect that this is the case.

3. Variables associated with the administration of questions

Pretests typically exist in several versions that may feature the same questions in a different order, or in combination with additional questions. Because performance on a given question might be affected by those that precede it, which could serve as hints or distractors, different versions may tend to elicit higher or lower numbers of correct answers. The different versions generally result from ongoing efforts to pinpoint difficulties and better prepare students for subsequent tutorial instruction. Therefore, it is also possible that before lecture and after lecture samples may differ with respect to which version or versions are represented. Therefore, version is used as a variable in the regression analysis.

A complication is that pretest participation is generally less than 100% (averaging 77%). We have found that higher performing students are more likely to take the pretests and thus participating students are not necessarily representative of the class as a whole. If the type of student who takes the pretest also varies according to whether instruction on the topic has been completed or not, then there could be a systematic difference between before and after classes. However, students only know the title of the tutorial prior to logging on to take the pretest. In most cases, the titles are sufficiently broad that they do not indicate the exact nature of the material that is relevant. For example, the questions about forces are associated with a tutorial simply called *Newton's Second and Third Laws*. We have no reason to suspect that students whose lectures have covered the second law but not the third would be inclined to skip the entire pretest as a result. Nevertheless, in the absence of information about how student characteristics affect their pretest participation, we can consider the percentage of enrolled students who took a given pretest. The higher the level of participation, the more representative the students will be. Therefore if there is a correlation between performance and the percentage who took the pretest, it may indicate that pretest participation is higher among a certain subset of students.

Most other factors associated with pretest administration are assumed to affect all classes more or less equally. For example, the fact that students do not earn points for the correctness of their responses may affect the degree to which the results correspond to those that would be obtained under graded conditions. The extent to which students avail themselves of other resources (textbooks, websites, other students) while they take the pretest may also affect the degree to which the results correspond to those that would be obtained under closed-book conditions.

B. Regression models

Regression models including some or all of the variables mentioned above were constructed for each question. Some variables are mutually correlated and therefore were not

included in the same model to mitigate the effects of multicollinearity. For example, course schedules and room assignments ensure that some classes are only offered at certain times of day in a given quarter (e.g., a section of Mechanics is held at 2:30 pm in spring quarter, but not in autumn or winter). These schedules have not changed substantially in more than a decade. Because of this association, a variable referred to as “schedule,” which combines academic quarter and scheduled lecture time (e.g., autumn quarter, 9:30 am) was used in some models and compared to the inclusion of either time or quarter. Also, because some versions of a given pretest are used for an extended period then replaced with another, version and academic year are correlated, but not as strongly as quarter and time of day. Nevertheless, models were tried that included version, academic year, and both.

Categorical variables (schedule, version, quarter, time of day, and occurrence of instruction) were converted to *dummy variables*. For example, “autumn 9:30” is a variable for which a class has a value of “1” if lectures took place then, but “0” otherwise. If there are n possible combinations of quarter and time, $n - 1$ dummy variables are needed. Thus the regression equation indicates the expected value of performance relative to that associated with the n th level.

The continuous variables class size and percent who answered the question are *centered* about their respective averages to ensure that the intercept of the regression equation will not refer to a hypothetical class of 0 students of whom 0% answered the question. Academic year is adjusted to 1999–2000 (the earliest year for which data are included) as “year 0,” and thus the intercept of the regression equation will not refer to a class that took place in year 0 BC.

A linear model of dependence is assumed for all of these variables for simplicity. However, some of them are clearly going to have nonlinear effects at the limits. For example, the percentage who answered the question has clear upper and lower bounds. However, within a certain range a linear model is unlikely to change whether or not a relationship exists, assuming that any regression coefficients are not taken too seriously.

The most complete regression equation for performance on a given question is

$$\begin{aligned} \pi = & \beta_0 + \beta_I I + \beta_C C + \beta_P P + \beta_A A + \sum_{j=1}^{n-1} \beta_{V_j} V_j \\ & + \sum_{j=1}^{m-1} \beta_{Q_j} Q_j + \sum_{j=1}^{p-1} \beta_{T_j} T_j + \sum_{j=1}^{q-1} \beta_{S_j} S_j + \epsilon_i. \end{aligned}$$

The variables in this equation are

- I , which is a dummy variable with a value of 0 if lecture instruction had not occurred, and 1 if it had,

- C , which is the number of students above or below the average class size,
- P , which is the percentage of enrolled students who answered the question above or below the average,
- A , which is the number of academic years elapsed since 1999–2000,
- $\{V_j\}$, which are dummy variables representing different versions,
- $\{Q_j\}$, which are dummy variables representing different academic quarters,
- $\{T_j\}$, which are dummy variables representing different times of day at which lectures took place,
- $\{S_j\}$, which are dummy variables representing different schedules (combination of quarter and time),
- ϵ_i , which is the “error” and represents the contribution of variables not included.

The coefficients, which are unique to a specific question, can be interpreted as follows:

- β is the intercept, and gives the expected performance of a class of average size, of which an average percentage of students answered the question, that took place in academic year 1999–2000, that had version 1 of the pretest on which the question appeared, that took place at 9:30 am in autumn quarter and that did not receive instruction.
- β_I gives the expected increase in performance if lecture instruction occurred.
- β_C gives the expected increase in performance for every additional student enrolled.
- β_P gives the expected increase in performance for every additional 1% of the class that answered the question.
- β_A gives the expected increase in performance for every academic year elapsed since 1999–2000.
- β_{V_j} gives the expected increase in performance if the question appeared on version j , rather than version 1.
- β_{Q_j} gives the expected increase in performance if the lecture took place in quarter j rather than autumn quarter.
- β_{T_j} gives the expected increase in performance if the lecture took place at time j rather than at 9:30 am.
- β_{S_j} gives the expected increase in performance if the lecture took place at schedule j rather than at 9:30 am in autumn quarter.

(Note that if either time or quarter is included, then schedule is not).

The significance testing described earlier was essentially an attempt to fit the data to a linear regression model with a single parameter, β_I . Clearly, adding parameters to a model will increase its apparent fit to the data. Therefore models that incorporate some of all of the additional variables identified above will appear to fit the data better than the original single-parameter model, as reflected in a higher coefficient of determination, R^2 . However, not every increase in R^2 is significant. To assess whether adding

parameters results in a genuine improvement, one can compare the *adjusted* coefficients of determination (“adjusted R^2 ”) for the new and original models. The adjusted R^2 takes into account the degrees of freedom in the model and does not necessarily increase with an increase in the number of parameters. The model that maximizes the adjusted R^2 can be considered the most parsimonious. For example, if adding “time of day” to a model that includes only “instruction” does not increase adjusted R^2 , then we can assume that the more complete model does not in fact fit the data better than the simpler one, even though R^2 itself will have increased. More robust techniques exist for comparing models, e.g., applying an F test to the sum of the squares of the residuals, and these might be appropriate if precise estimates of model parameters were important. However, in this case, the primary goal is simply to assess whether models with some or all of the variables described above support the conclusions of the earlier significance testing. In those cases in which the conclusions differ, the adjusted R^2 values are provided.

C. Results

The results are summarized in Table III. The first column repeats results from Table II. The second and third columns provide information about the best regression model for

each question. The (nonadjusted) coefficient of determination (R^2) indicates how much of the variance is explained by the model. The value p_{model} indicates the probability that *all* of the coefficients are zero; a low p_{model} suggests the model is at least somewhat successful in fitting the data. The next column gives the estimated coefficients for the variable instruction, with an indication of their level of significance. The leftmost column lists other variables that were significant in the best model.

As shown in the Table, for 16 of the 20 questions, the results are essentially consistent with the earlier findings, at least insofar as determining whether lecture instruction is a significant explanatory variable. Competing models tended to differ with respect to the magnitude of the coefficient β_1 and its level of significance, but the difference was one of degree. For example, on question 12, simple tests showed a statistically significant difference of -0.05 between the before lecture and after lecture samples ($p < 5\%$). Regression analysis confirms that lecture instruction is a significant (negative) variable ($\beta_1 = -0.05$, $p < 5\%$). For question 3, the estimated coefficient for instruction is both smaller and less significant ($\beta_1 = 0.09$, $p < 1\%$ in the simplest model vs $\beta_1 = 0.01$, $p < 10\%$ in the best model) but according to criteria established earlier, both count as evidence of the impact of lecture instruction.

TABLE III. Results from regression analysis.

	Results from significance testing (repeated from Table II)		Results from best regression model		
	Difference	Significance of entire model	Coefficient of determination	Estimated coefficient for instruction	Other significant variables
	$\pi_{\text{after}} - \pi_{\text{before}}$	π_{model}	R^2	β_1	
1	0.01	<1%	0.24	-0.03	Academic year, class size, schedule
2	0.07***	<1%	0.42	0.03	Academic year, class size
3	0.09***	<1%	0.43	0.01*	Academic year, class size
4	0	>10%	0.05	-0.02	
5	0	<1%	0.24	-0.02	% took
6	0.08***	<1%	0.46	0.05**	% took
7	-0.04**	<10%	0.21	-0.06***	
8	-0.01	>10%	0.20	-0.03*	
9	0.14***	<1%	0.41	0.04	Academic year
10	0.37***	<1%	0.84	0.45***	Academic year, schedule
11	0.07**	<5%	0.30	0.8**	Schedule
12	-0.05**	<1%	0.28	-0.05*	Academic year, % took
13	0.01	<5%	0.18	0.02	Academic year
14	0.04***	<5%	0.17	0.04**	Class size, schedule
15	0.12***	<1%	0.37	0.09***	Academic year
16	0	<5%	0.25	0	Class size, schedule
17	-0.02	<1%	0.55	-0.02*	Academic year, schedule
18	0.05*	<1%	0.55	0.04*	Version
19	0.08***	<1%	0.21	0.08***	
20	0.1***	<5%	0.19	0.09**	Class size, schedule

***, ** and * indicate significance at the 1%, 5%, and 10% levels, respectively.

FIG. 3. Estimated coefficients for the parameter instruction in the regression models (β_1), arranged in the same order as in Fig. 2. Bars shaded dark gray, medium gray, and light gray indicate effects significant at the 1%, 5%, and 10% levels, respectively.

For two cases (questions 2 and 9) in which simple statistical tests indicated a significant difference between before lecture and after lecture samples, regression analysis favored other explanatory variables. For question 2, which had a difference in average performance of 0.07 ($p < 1\%$), the coefficient for instruction was not found to be significant ($\beta_1 = 0.04$, $p > 15\%$) while academic year and class size had significant effects. The best regression model in this case had an adjusted R^2 of 0.34 ($p_{\text{model}} < 1\%$), compared to 0.11 for the simplest model ($p_{\text{model}} < 5\%$).

Similarly, for question 9, which had a difference in average performance of 0.14 ($p < 1\%$), the coefficient for instruction was not significant ($\beta_1 = 0.04$, $p > 25\%$) while academic year had, again, a significant effect. The best regression model in this case had an adjusted R^2 of 0.34 ($p_{\text{model}} < 1\%$), compared to 0.15 in the simplest model ($p_{\text{model}} < 1\%$). This question stands out because of the large discrepancy. However, there are very few classes in the after lecture sample and they are all from the same year. Therefore the results for this case must be viewed as more tentative than the others.

For two cases (questions 8 and 17) in which simple tests showed no statistically significant difference between before lecture and after lecture samples, regression analysis indicated that the occurrence of instruction was in fact (just) significant. For question 8, simple tests indicated the difference of -0.01 was not significant, but regression analysis estimated the corresponding coefficient to be -0.03 ($p < 10\%$). For the best model, the adjusted R^2 was 0.1 ($p_{\text{model}} > 10\%$); for the simplest, -0.005 ($p_{\text{model}} > 35\%$). Thus, while one model appears superior to the other, neither one can be considered a good fit. For question 17, simple tests indicated a nonsignificant difference of -0.02 ($p > 10\%$), while regression analysis estimated the coefficient for instruction to be -0.02 ($p < 10\%$). For the best model,

the adjusted R^2 was 0.46 ($p_{\text{model}} < 1\%$); for the simplest, 0.004 ($p_{\text{model}} > 25\%$).

For nine questions, the best model included the variable academic year. In two models the coefficients for this variable were negative, but in the others it was positive, ranging from 0.003 to 0.016. In other words, for each year that elapsed, an additional 0.3% to 1.6% of students would answer correctly, even controlling for other variables. Over a period of ten or more years, this effect was substantial. For seven models, some sort of information about the class schedule was significant but there was no overall trend in favor of one quarter or time of day over another.

The results are also presented graphically in Fig. 3. The bars represent the coefficients for the variable lecture instruction β_1 and are arranged in the same order as in Fig. 2. They are shaded according to their significance in the model, with the dark, medium, and light gray bars indicating significance at the 1%, 5%, and 10% levels, respectively. As can be seen by comparing Figs. 2 and 3, there is reasonable agreement about the role of lecture instruction in predicting performance. However, the regression coefficients tend to be smaller and less significant.

VI. DISCUSSION

A. Trends in the results

As shown, the impact of lecture instruction differed for different questions. These results raise an important question: why is performance on some questions more easily influenced than others? The question that stands out for having a large, positive difference is No. 10, which concerns electric flux and is arguably one of the more formal and abstract of the set examined here. For several other questions in this study, knowledge of formal concepts is not strictly necessary; for example, predicting which of two dropped spools will strike the ground first (question 8).

Others deal with formal concepts that also have everyday meanings or that students are likely to have encountered in previous instruction; for example, describing the direction of the acceleration of an object at the peak of its trajectory (question 1). While students may not apply formal definitions and procedures in responding to such questions, it is rare for them to claim ignorance of the concepts. In contrast, question 10 refers to a concept for which one might expect such a reaction and, consequently, a clear difference in pre- and postlecture instruction responses, as was the case. Questions 1, 8, and 10 seem to follow a pattern in which benefits of instruction emerge predictably as the formality of the questions or concepts increases. However, questions on electric field (No. 13), charge density (No. 14), and phase difference (No. 18) do not easily fit this pattern. Further research is needed to identify the factors that make performance on some questions more easily influenced by instruction than others.

Another notable feature of the data is the presence of questions (7, 8, 12, and 17) on which performance is apparently negatively affected by relevant lecture instruction. Other studies have documented temporary decreases in performance during instruction [9], or “U-shaped” learning [10]. For questions 7, 8, 12, and 17, the data were collected up to 12 days after lecture instruction. No obvious time dependence is suggested by the data, but there are too few data points for a pattern to be reliably detected. The permanence of the effect is a subject for further investigation, as is the significance of the type of question on which this effect was observed.

B. Commentary on instructional significance

The focus in this article has been on statistically significant differences. However, statistical significance by itself is not sufficient to guide practical decisions. Even effect sizes only provide a sense of how large an effect is in relation to the width of the underlying distributions, not the degree to which a difference would be noticeable to students or instructors. The decision as to whether the benefits of an intervention warrant its costs is complex and requires familiarity with the context in which data are obtained. Thus at UW we tend to focus on instructionally significant effects: those that are “real”—not the results of random variation, plausibly attributable to the intervention in question—not some other factor, and substantial enough to at least warrant consideration of the intervention in light of the costs associated with its implementation.

In a randomized control trial, straightforward statistical testing can determine whether an effect is real and, because of the controls in place, associate it directly with the intervention. However, randomized control trials, while conceptually simple, are not the only way to test the effectiveness of an intervention. In situations in which randomization raises practical and/or ethical issues,

nonrandomized “quasiexperiments” can provide convincing evidence [11]. A key challenge for such experiments is attributing an effect to the intervention in the presence of many uncontrolled variables. In the data set discussed here many such variables were found to be negligible for determining whether an instructional intervention (in this case, lecture instruction) has an effect. Specifically, controlling these variables through multiple regression was unnecessary for detecting differences, in most cases. Nevertheless, accumulating additional data helps increase confidence and can indicate whether an effect will be present on average across many different conditions. A focus on differences substantial enough to have implications for instruction also mitigates the effect of the lack of controls. That is, if the effect due to an intervention is easily obscured by the natural variations likely to be found in real classrooms, it may not have much practical importance.

At UW, tutorials have been implemented in hundreds of classes over more than 25 years. Experience with pretest results suggests that a difference in performance of 10% or more—even between individual classes—is *likely* to be real, related to an intervention, and in the realm in which considering a change to instruction is starting to be reasonable. In other words, a difference of less than about 10% is not typically considered to have strong implications for instruction even if it proved to be statistically significant. Some data from this study provide an illustration.

For the 13 questions that exhibited a statistically significant difference in average performance (at least according to simple tests), the standard deviations for the pooled data range from 0.05 to 0.14 with a median of 0.09. The differences between the means of the two samples range in absolute value from 0.04 to 0.37. All differences of 0.10 or more proved significant at the 1% level with effect sizes ranging from 0.7 to 5.9. All differences between 0.05 and 0.10 also proved significant, but at differing levels. There are also some cases (e.g., question 14) in which even a small difference in mean performance proved to be statistically significant. However, such a difference may not be instructionally significant. That is, the time and effort spent on instruction may not be well spent. While these examples refer to instruction in a lecture hall, the same holds true for interventions like tutorials. In a situation in which tutorials are already in use, and one can simply replace one worksheet with another, the cost of the intervention might be sufficiently low that the change is justified. However, as the time and resources required to implement the change increase, the less likely it is to be warranted.

C. Study limitations

In interpreting the results presented here it is important to acknowledge that a number of potentially important variables were not taken into account. Therefore it is not surprising that in all but three cases, a majority of the variance is unexplained (i.e., R^2 is below 0.5). Foremost

among the neglected variables is the quality of lecture instruction. The classes included in this study were taught by almost 100 individual faculty members who represent a wide range of style, experience, and perceived effectiveness, as measured by student evaluations of teaching. These instructors—while following a common syllabus—prepare their own lectures and choose their own sample problems, derivations, and lecture demonstrations. Recently most have also used “clicker” questions of their own design. Thus the details of exactly what occurred in a given lecture varied considerably from class to class. However, we have no reason to expect a *systematic* association between lecturer expertise and the timing of the pretest questions. Therefore, while it is certainly possible (even probable) that some features of lecture instruction affect performance, this variation is not expected to alter the results reported here, which address some sort of “average” instruction. Moreover, with a median standard deviation of 0.08, the upper ends of many after lecture distributions do not substantially differ from the lower ends. If we assume the upper ends are due to superior instruction, then the effects thereof may be relatively small.

It is also possible that some of the unexplained variance reflects differences in the time elapsed between lecture instruction and administration of the pretest. The ability of students to answer correctly may depend on how recently the topic was discussed and what has happened since, e.g., reinforcement and application through additional instruction, homework problems, etc. For some of the questions discussed here, performance appears to be somewhat better if a few days have elapsed; for other questions there is no clear trend. Additional analyses are planned to examine this.

Also, as noted earlier, the introduction of a topic in lecture might not have occurred when scheduled according to the syllabus. The most likely scenario is that actual instruction lagged behind the schedule, so that a class assigned to the after lecture sample might really belong in the before sample. If these mislabeled classes perform at lower levels than those in which instruction had actually taken place, then average after lecture performance would be artificially lowered. The effects of such a scenario can be estimated. For example, if we assume that the poorest performing 10% of each after sample had been mislabeled and repeat the analysis with “corrected” labels, then 14 out of the 20 questions would exhibit a significant positive effect of lecture instruction (instead of 11), two would exhibit a significant negative effect, and four would exhibit no effect (instead of seven). Thus, there would still be some variation in the sign and magnitude of the impact of lecture instruction. However, it is more likely that mislabeling would occur at random, not in such a systematic fashion. Therefore even if 10% of the classes were mislabeled, the effects are likely to be smaller than those illustrated in this hypothetical scenario. It is also important to note that

instruction in lecture is unlikely to lag behind the schedule by more than a day or two; therefore, the longer the elapsed time, the more reliable the data. Attempts to gauge the effect of elapsed time on performance should also help mitigate this effect.

Finally, it is important to note that the questions discussed here are not all independent (some appear together on a single pretest), nor are they necessarily representative of qualitative questions in general. Therefore some quantitative patterns, such as the fraction of questions that exhibit an instruction-related decrease in performance, are not meaningful outside this study.

VII. CONCLUSION

The results presented here suggest that the effect of lecture instruction on performance can vary considerably. While regression models differed in the extent to which instruction was linked to an improvement in performance, they tended to support the results of simple statistical tests, and, for that matter, “naked eye” examination of the data.

For those questions on which a detectable improvement was seen, the effects varied in magnitude. Average post-lecture performance also varied over a wide range. In some cases, an average of around 70% of students answered correctly. In others, while a noticeable improvement could be attributed to lecture instruction, final performance did not reach a very high level. In almost all cases, there was a significant overlap between prelecture and postlecture results, indicating that some classes began at a higher level than others reached, even after relevant instruction has been completed. Thus, for most questions the “value added” by lecture instruction is comparable in magnitude to that of other factors.

This article also presents robust evidence that for some questions, lecture instruction has no discernible effect. While it is possible that comparable numbers of classes experienced positive and negative effects, it could be that essentially none of the classes experienced an effect.

Evidence was also presented that lecture instruction can be associated with a decrease in student ability to answer certain questions correctly. Whether the effect is temporary or permanent is not known.

It is important to emphasize that the results presented here do not imply that direct attempts to address the pertinent concepts in lecture will be met with similar levels of success. Lecture instruction might very well be effective at improving performance on *all* of the tasks discussed here if the relevant concepts are targeted specifically. That is, the mere continued presence of incorrect responses after instruction does not imply the presence of underlying difficulties that are resistant to change, a judgment that requires evidence that concerted efforts have met with limited success. Thus these results do not indicate that lecture instruction *cannot* increase performance on these topics or questions.

It is also important to again emphasize that performance, defined here as the percentage of students who answer correctly, is only one measure of understanding. It is possible that while the overall percentage of correct answers does not change substantially, there could be shifts in the frequencies of various alternative answers, and/or of the soundness of the reasoning that led to correct answers. The results in this paper are mute about these possibilities.

Finally, while performance on some questions did prove to be amenable to instruction, there is no clear pattern that

differentiates these questions from those on which no effect (or a negative effect) was seen. Explaining *why* performance on some questions or topics is more easily influenced by lecture instruction is an area for further investigation.

ACKNOWLEDGMENTS

The author appreciates the contributions of other members of the Physics Education Group, the suggestions of anonymous reviewers, and the support of the NSF through Grant DUE 0088840.

-
- [1] For early examples, see L. C. McDermott, Millikan Lecture 1990: What we teach, and what is learned—Closing the gap, *Am. J. Phys.* **59**, 301 (1991); R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *Am. J. Phys.* **66**, 64 (1998).
 - [2] A comprehensive review can be found in D. E. Meltzer and R. K. Thornton, Resource Letter ALIP-1: Active-learning instruction in physics, *Am. J. Phys.* **80**, 478 (2012).
 - [3] L. C. McDermott and P. S. Shaffer, and the PEG at UW, *Tutorials in Introductory Physics* (Pearson, Upper Saddle River, NJ, 2002).
 - [4] A good overview of randomization tests can be found in P. Dugard, P. File, and J. Todman, *Single-Case and Small-n Experimental Designs: A Practical Guide to Randomization Tests* (Routledge, New York, 2011).
 - [5] For question 14, excluding a clear outlier changed the verdict from “no difference” to “after lecture” is better than “before.” This outlier is excluded from the remaining analysis for questions 12 to 14, which share the same data set.
 - [6] U.S. Department of Education, Translating the Statistical Representation of the Effects of Education Interventions Into More Readily Interpretable Forms, NCSE 2013-2000, November 2012.
 - [7] A discussion of the impact of variations in student characteristics can be found in R. Theobald and S. Freeman, Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate STEM education research, *CBE Life Sci. Educ.* **13**, 41 (2014).
 - [8] See <http://opb.washington.edu/sites/default/files/opb/Policy/UW%20Academically%20Elite%20-%20Not%20Economically%20Selective.pdf>; <http://admit.washington.edu/QuickFacts>.
 - [9] See, for example, E. C. Sayre, S. V. Franklin, S. Dymek, J. Clark, and Y. Sun, Learning, retention, and forgetting of Newton’s third law throughout university physics, *Phys. Rev. ST Phys. Educ. Res.* **8**, 010116 (2012).
 - [10] A review of U-shaped learning in the context of physics can be found in P. J. Camp, U-shaped development of Newtonian concepts, [arXiv:1007.1647](https://arxiv.org/abs/1007.1647).
 - [11] The relationship between randomized control trials and quasiexperiments is discussed in W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Houghton-Mifflin, Boston, 2001).